ARI Research Note 91-19

Conceptual Models of Unit Performance

**Stuart H. Rakoff, Kathryn B. Laskey,
F. Freeman Marvin, and Jeffrey S. Mandel**

Decision Science Consortium, Inc.

for

**Contracting Officer's Representative
James H. Banks**

**Presidio of Monterey Field Unit, California
Howard H. McFann, Chief**

**Training Research Laboratory
Jack H. Hiller, Director**

January 1991

**United States Army
Research Institute for the Behavioral and Social Sciences**

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

Decision Science Consortium, Inc.

Technical review by

James H. Banks

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 91-19
6a. NAME OF PERFORMING ORGANIZATION Decision Science Consortium, Inc. 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute Presidio of Monterey Field Unit
6c. ADDRESS (City, State, and ZIP Code) 1895 Preston White Drive, #300 Reston, VA 22091	7b. ADDRESS (City, State, and ZIP Code) P.O. Box 5787 Presidio Monterey, CA 93944-5011
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) --	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-89-C-0660
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 65502M; PROJECT NO.: 770; TASK NO.: N/A; WORK UNIT ACCESSION NO.: N/A
11. TITLE (Include Security Classification) Conceptual Models of Unit Performance	
12. PERSONAL AUTHOR(S) Rakoff, Stuart H.; Laskey, Kathryn B.; Marvin, Freeman; and Mandel, Jeffrey S.	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM 89/03 TO 89/08	14. DATE OF REPORT (Year, Month, Day) 1991, January 15. PAGE COUNT 48
16. SUPPLEMENTARY NOTATION James H. Banks, Contracting Officer's Representative	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Neural network models; AirLand Battle; Unit performance; Training measurement; Army doctrine; SBIR

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report summarizes work testing the usefulness of neural-network models for measuring and predicting Army unit performance in settings such as the National Training Center. A back-propagation neural-network model was developed and trained from wargaming simulation. Results suggest that such a model can train to recognize unit success and failure in simulated engagements. Further work will require access to clean, large, data sets, such as those available from SIMNET. In addition, an expert-based preprocessor is suggested as a useful approach to implementing a model.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL James H. Banks	22b. TELEPHONE (Include Area Code) (408) 647-5482 22c. OFFICE SYMBOL PERI-IOC

DD Form 1473, JUN 86 *Previous editions are obsolete.*

CONCEPTUAL MODELS OF UNIT PERFORMANCE

EXECUTIVE SUMMARY

Requirement:

To investigate the usefulness and potential of natural-analogy and neural-network models as methods for measuring and predicting unit performance in training settings such as the National Training Center (NTC)

Procedure:

Literature on neural networks and the doctrine of unit mission in combined-arms task forces was researched, and a number of data sources were investigated as possible sources of data for building and validating a prototype model. Data from the NTC, hypothetical data, and data collected from repeated plays of a wargaming model were used to construct and solve neural-net models.

Findings:

Limited resources prevented new data collection, and the quality of the available data limited the generality of the results of this effort. A back-propagation model was built and trained from the wargaming data and was able to outperform human experts in predicting the outcome of simulated engagements. Concepts that appear prominent from this neural-network model are closely comparable to major tenets of Army doctrine.

Utilization of Findings:

Alternative data sources will be required to extend this research to a more useful application stage. Investigation reveals that the Army/DARPA SIMNET system, a combined-arms simulation-training system, is capable of producing a data set comparable to that from the NTC. A recommended next step is to collect engagement data from SIMNET as a means of providing enough quality data to train a neural-net model of unit performance. In addition, follow-on research that incorporates an expert judgment-based preprocessor to a neural net is recommended as the path most likely to produce a model of unit performance that can be used to evaluate NTC activities.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CONCEPTUAL MODELS OF UNIT PERFORMANCE

INTRODUCTION

As the Army enters a period during which budget and manpower resources are likely to be constrained, the effective allocation of resources among competing programs and their justification within DoD and to the Congress becomes increasingly critical. Effective allocation of these scarce resources can be enhanced if the outputs produced from those resources can be measured and predicted, allowing the Army to formulate decision rules which specify the most efficient resource allocations. The problem is complicated because the most important Army organizational output--unit performance in combat--cannot be easily conceptualized or measured in peacetime. Without a valid and usable measure of unit performance, it is not possible to evaluate the effectiveness of resource inputs, build a composite measure of organizational outputs nor link them together into an integrated cost-effectiveness measurement system. Additionally, robust measures of unit performance are important diagnostic tools for evaluating and improving unit training.

The Army has tried several approaches to this problem. One approach is to use readiness as a proxy for performance. Readiness is an inherent, peacetime quality which every unit possesses to some degree. Common measures of readiness include equipment and personnel status, skill qualification test results, and reenlistment rates in the unit. The Army currently uses a number of readiness reports and indicators, including the Unit Status Report (USR) submitted monthly. Many questions have been raised concerning the USR (Shishko and Paulson, 1981), particularly because the report lacks measures in the particular kinds of missions the unit can undertake, or fails to answer the question "Ready for what?"

In recent years the Army has begun to develop measures of unit performance which go beyond the simple and static basis of the USR. These systems have been focused on identifying the tasks that the unit is expected to perform and its ability to accomplish these tasks. Much effort has been put into developing these standards of unit performance. Traditionally, the standards have involved a list of tasks at various degrees of detail, which, if accomplished one at a time, will define successful unit performance. The most common of these lists are the Army Training and Evaluation Program (ARTEP) manuals, now known as Mission Training Plans (MTP). Of course, accomplishment of the task lists cannot assure a successful *outcome*, but only desirable unit *output*. The specific mission, enemy forces, terrain, and other factors also affect the combat outcome. Most units undergo an ARTEP evaluation every 18 months, and spend much of the time between ARTEPs practicing the tasks which will be tested.

Even more thorough performance measurement takes place at the Army instrumented training ranges such as the National Training Center at Fort Irwin, California, or the Joint Readiness Training Center at Fort Chaffee, Arkansas. At these sites (and additional facilities to be developed) combined arms task forces are able to maneuver and train against a skilled resident opposing force in what has become the most realistic (and expensive) replication of real combat available.

Finally, new technology, as represented in the SIMNET system, may eventually allow computer simulations of combined arms forces as both training and performance evaluation devices. If successful, these networked simulation systems will be cheaper and more accessible than the currently limited opportunities for units to move to an instrumented and monitored location for realistic training and evaluation.

What all of these measurement systems have in common is that they are basing their evaluation of performance on a series (sometimes very large) of discreet individual and collective tasks for a particular unit. Typically, each task is scored on a go/no go basis. The measure of overall unit capability or performance is then built from the ability of the unit to achieve each of the subsidiary goals or meet lower-level criteria. For example, overall readiness may combine measures such as the percentage of the unit's vehicles or systems which are ready, or the number of skilled and trained soldiers available compared to the required number on the unit's manpower document. Overall performance may combine measures such as whether or not the unit used proper procedures during movement to contact, or reacted to an NBC alert correctly.

Because each of these sources of peacetime readiness or training performance information measures different qualities of the unit, such as potential firepower, or state of training and morale, a series of inferences is needed to develop their relationship to combat performance. The inference rules which allow these individual observable measures to be combined into a measure of overall unit performance are highly subjective, and, therefore, are controversial. Nevertheless, these inferences provide the critical link between peacetime readiness, training proficiency, and expected unit performance in combat.

In addition to the inference difficulties, these task lists are expensive to develop and use. They are typically developed by a work-breakdown analysis of simulated combat, resulting in hundreds or thousands of separate tasks. These tasks are themselves complex and often ambiguous, complicating the measurement problem (Kahan, et al., 1985; Shisko and Paulson, 1981). Validity is achieved by observing and interpreting field exercise data from the National Training Center (NTC) and other training areas.[1] Unit performance conceptualized this way is also costly to apply because it requires a large number of "observer/controllers" to record the lower-level tasks as the unit performs them. Evaluation of overall unit performance then requires a complex synthesis of performance of individual

---

[1]For example, the study being conducted by BDM, Inc. for ARI is attempting to develop task-based performance measures for task forces at the National Training Center. See Forsythe, Thomas, "A Research Concept for Developing and Applying Methods for Measurement and Interpretation of Unit Performance at the National Training Center," ARI Field Unit at Presidio of Monterey, California, January 1987.

tasks. While this methodology provides some diagnostic capability, the costs, in many cases, far exceed the usefulness.[2]

Measures of unit performance which are less subjective and more efficient would improve the Army's ability to assess peacetime readiness, allocate resources, and diagnose training needs. The inspiration behind this project is to apply to the measurement of unit performance a new, "natural-analogy" approach to modeling complexity.

A unit is a complex system. Its performance depends on many component processes which interact to produce macro-level behavior that is not always precisely predictable, even when the micro-level processes are well understood. There has been a recent surge of interest in modeling complexity in natural and organizational processes. In disciplines as diverse as physics, economics, and biology, it is being realized that traditional reductionistic models, based on a mechanistic model of the universe, may not be sufficient to characterize global, "emergent" properties of complex systems. With the advent of more powerful computing technology, it is becoming possible to simulate complex systems of interacting subunits, and study how micro-level assumptions on the interactions among component parts influence behavior of the systems at a macro level. The impetus for many models of complex systems in a number of domains has come from attempts to describe the behavior of complex systems arising in nature. For example, nonlinear dynamics and chaos theory (cf., Gleick, 1987) originally arose out of the physics of turbulent systems, but is being applied to model social, biological, and economic systems as well. Genetic algorithms (Holland, 1975) apply insights from the theory of evolution to general optimization problems. Neural-network models were developed as models of cognition in the brain, and are being applied to problems including optimization, vision, natural-language understanding, and a host of others (cf., DARPA, 1988).

The goal of this project is to explore the degree to which this new, "natural analogy" approach to modeling complexity is useful for modeling and measuring unit performance. We have chosen to focus on the neural-network approach for Phase I. The basic hypothesis is that a neural-network model of unit performance could take account of the multiple aspects of tactical unit performance as would be observed by a military expert. Such a model could address a fundamental aspect of unit performance that cannot be addressed by the current "determinants" paradigm: the complex interaction of unit components as they work together to perform the unit's mission. The research effort determined the sources and meaning of available data and proposed and tested an alternative inferential relationship linking these indicators to unit performance. Specifically, the technical objectives of Phase I were:

---

[2]For an excellent discussion of the role of the observer/controllers at the National Training Center, see the outbriefing presented by COL Larry E. Word, Former Director of the Joint Readiness Training Center and reported in Word, 1987.

- to identify the multiple indicators that may be used to estimate unit performance;

- to develop a methodology for combining them into a single measure of performance; and

- to design and test a prototype of the methodology.

Key questions that were addressed during Phase I were:

- Can a methodology be developed that captures the complexity of the problem yet is simple enough to be used by a variety of Army users?

- How should the methodology be implemented?

- What is needed to develop a complete model during Phase II?

This Phase I effort has determined the technical feasibility of modeling unit performance using natural analogy, specifically neural networks trained by back propagation. This approach consisted of two tasks: (1) identification of performance factors and development of a methodology for combining these factors into a single model of unit performance; and (2) the design and validation of a prototype model for estimating unit performance.

CONCEPTUAL MODELS OF UNIT PERFORMANCE

The music is not in the notes; it's in the pauses between the notes.

Approach

Defining and validating a measure of unit performance is essentially a problem of arriving at an appropriate measurement model for overall unit performance. Judging that one unit i "better" than another unit on a complex set of factors is not that difficult for the human; Army leaders make these judgments every day in the performance of their command duties. Although these commanders may not easily be able to specify precisely what factors lead them to classify units as good or poor, and although their judgments may be incorrect (that is, lack validity), commanders interviewed in connection with other research have no difficulty articulating a level of unit effectiveness and quality. The approach to the problem of developing a more rigorous and analytic measurement model of unit performance begins, therefore, with the assumption that military "experts" can recognize good units when they observe them.

Commanders' assessments of unit performance are based upon a variety of data, both "hard" and intangible. Inventories of personnel and equipment available compared to the number required provide one set of "hard" measures, as do scores on soldier skill tests. Measures of soldier quality also contribute to unit performance, although these apparently hard data are more difficult to use because the evaluator must define a more complex inference structure linking AFQT scores to performance. Intangible data such as the assessment of unit morale or training level, and even to some extent performance on exercises such as unit ARTEPs, are even more difficult because easily applied metrics for these soft factors do not exist; individual judgments become crucial here.

Those considerations apply to the problem of making a static assessment of overall unit quality, but in this research the objective was to develop methods and models for estimating unit performance directly in a combat engagement, or at least in a simulation of that engagement at the National Training Center. In this case, the problem is to predict how well a unit will accomplish a combat mission such as a defense in position, withdrawal, or deliberate attack. Moving the focus of the performance measurement problem to actual (or simulated) combat engagements changes the nature of the measurement problem substantially.

Much of the focus of both military historians and operations researchers has been on developing analytic or mathematical models of the combat engagement. Typically these force-on-force models concentrate on variables such as force ratios, firepower and casualty rates. In the typical combat simulation model (such as the JANUS model used by the Army Concepts Analysis Agency for Total Army Analysis), exchanges between individual weapons systems are played and aggregated to arrive at force-attrition ratios which determine engagement and battle outcomes. But many critics of these approaches argue that these stochastic models fail to

capture the dynamics or human dimension of actual combat.[3] Some of these variables can in fact be quantified, although the precise methods and data are controversial.[4] But an alternative approach based on a more holistic view of the problem is appropriate as well.

Operations-research approaches to these measurement problems, as well as most of the existing performance-evaluation methodologies currently in place, are based upon modeling or measuring very micro-level phenomena or tasks. This approach has been selected largely because of its diagnostic value as a training aid. Thus, the ARTEP evaluation system is based on the accomplishment of hundreds of specific discreet tasks; simulation models may replicate thousands of system vs. system engagements. Current evaluations of performance at the NTC are based upon completion of detailed checklists of tasks and activities relevant to a specific unit or element in each mission area. These discreet results are then aggregated to arrive at overall performance or predictions of outcomes.

But imagine an experienced officer standing on a hill overlooking an engagement about to take place. Our experience suggests (and we offer as a hypothesis) that this expert observer can tell from relatively straightforward observation of the forces moving to contact and the initial contact, what the outcome of the engagement will be; that is, he can predict how well the unit or task force will perform. He does this by recognizing patterns in the movement of vehicles and troop units, assessing the ways in which each side is taking advantage of METT-T, and applying his experience and expertise to process literally thousands of pieces of data. This processing is typical of the way in which humans are able to process vast amounts of data, and is quite different from the more linear types of models and problem-solving techniques common in analytic studies of military phenomena. This natural-analogy approach provides the basis for the research work reported here.

Measures of Unit Performance

Army doctrinal publications provide the starting point for understanding the macro factors which a natural-analogy model of unit performance will be designed to emulate. As will be discussed in more detail below, one major characteristic of this family of models is that it is able to recognize patterns of individual inputs in ways similar to the manner in which humans are able to recognize patterns in visual inputs. One major application of neural nets has, in fact, been in visual pattern recognition--nets have been taught to recognize letters of the alphabet

---

[3]The Military Operations Research Society (MORS) has in the past two years been sponsoring a working group dedicated specifically to the problem of improving the conceptualization and measurement of human-performance variables in these combat models.

[4]See the work of COL Trevor Dupuy for examples of attempts to quantify variables such as morale, leadership, and lethality.

from partial character representations and to recognize faces from photographs much as humans do. In these cases, the individual pixels of visual information are the inputs to the neural net, and the net model is then able to learn to put these individual pixel data together to recognize the patterns of pixels which represent characters of the alphabet, facial features, or terrain features.

Army doctrine also can be characterized as having broad concepts and patterns which are composed of more micro-data elements. At its broadest levels, Army doctrine (as presented in FM 100-5) offers eight major concepts which the Army believes will govern success on the battlefield of the future. These eight are, in turn, grouped into two major categories, entitled Dynamics of Combat Power and Airland Battle Doctrine. As these are discussed in the following sections, the reader should keep in mind the general nature of the concepts and think about how a trained observer would be able to recognize the presence or absence (or with even more difficulty, the amount) of each of these concepts in a real or simulated combat engagement.

Dynamics of combat power. "The dynamics of combat power decide the outcome of campaigns, major operations, battles and engagements. Combat power is the ability to fight. It measures the effect created by combining maneuver, firepower, protection and leadership in combat actions against an enemy in war" (FM 100-5, page 11).

This exhortation from Army doctrine illustrates the first of the two sets of high-level measures of unit performance. Maneuver, firepower, protection, and leadership are principles which are taught to Army leaders at all levels, and which form the core of successful unit performance. That is to say, a unit which under battlefield or simulated battlefield conditions is able to embody these principles in its actions will be successful. But the detailed task-list measures of unit performance currently in place do not appear to get at these overarching concepts successfully, although Army leaders are able to recognize their presence or absence. How then can these principles be captured in a unit-performance model?

"Maneuver is the movement of forces in relation to the enemy to secure or retain positional advantage. It is the dynamic element of combat--the means of concentrating forces at the critical point to achieve the surprise, psychological shock, physical momentum and moral dominance which enable smaller forces to defeat larger ones" (FM 100-5, page 12). The very terms in which maneuver is defined here defy measurement with task or check lists. In fact, the underlying concept of maneuver is highly relational--it concerns how Blue forces move with regard to how Red forces move. Thus, measurement of maneuver needs to be able to recognize the pattern of Blue force positions over time and in relation to objective, terrain, and Red forces.

Firepower implies much more than its common measurements, number of rounds fired, or tons of munitions delivered. The discussion of the firepower concept in Army doctrine highlights the same pattern and dynamic characteristics emphasized in the previous discussion of maneuver. Firepower must be measured in the context of what the Red Force is doing and what the Blue Force desires to prevent him from doing. Again, there-

fore, it is the pattern of interaction between Blue and Red forces which the measurement model must be able to capture. In the context of Airland Battle doctrine, firepower must be used to disrupt enemy maneuver, to destroy, delay or disrupt enemy forces, to damage or degrade enemy command and control and sustainment capabilities, and to overcome unfavorable force ratios. From the perspective of U.S. commanders, bringing firepower to bear effectively requires coordination of intelligence, force maneuver, command and control, and logistics functions. In short, measurement of the concept of firepower requires much more than simply counting shots fired, but requires extensive examination of the patterns of a number of other variables, each of which is difficult to measure in itself.

Protection, the third dynamic of combat power, "is the conservation of the fighting potential of a force so that it can be applied at the decisive time and place. Protection has two components. The first includes all actions that are taken to counter the enemy's firepower and maneuver by making soldiers, systems, and units difficult to locate, strike and destroy. ... The second component of protection includes actions to keep soldiers healthy and to maintain their fighting morale" (FM 100-5, page 13). Here again, the meaning of the concept extends well beyond what can be captured in a simple (or even extensive) task list. Judgments about the adequacy of protection activities require the evaluator to understand the whole complexity of the situation, to recognize patterns in the relationships between the elements of the Blue and Red forces, and to place all of these data in the context of METT-T. These are the kinds of judgments that trained human leaders and experts are able to make, but which have proven so difficult to capture in evaluation techniques and models.

"The most essential element of combat power is *competent* and *confident leadership*. Leadership provides purpose, direction, and motivation in combat" (FM 100-5, page 13). The complexity of leadership and its measurement is well documented in the literature. That literature suggests that leadership must be considered in the context of the situation and high levels of leadership often make the difference between success and failure, in both military and other settings. Leadership is also very difficult to measure directly--it is often adduced after the fact, rather than being observed during the combat engagement.

AirLand Battle doctrine. AirLand Battle is the Army's philosophy and concept for fighting a high-intensity war in the future, and in its doctrinal statement presents a high-level approach for organizing and utilizing combat forces in that war. As a statement of doctrine, AirLand Battle is necessarily vague and open to much interpretation, as a review of the military science and Army journals over the past 5 years will clearly demonstrate. In essence, AirLand Battle describes a modern battlefield which is intense, lethal, and dynamic, assumes a three-dimensional battlefield, and expects that operations on that battlefield will not be linear but will instead feature deep attacks to interject U.S. forces deep behind Warsaw Pact lines as well as similar Red incursions into deep areas behind US and NATO lines. The AirLand Battle will require commanders and soldiers to be smarter, plan and react faster, and use resources more decisively.

FM 100-5 outlines four basic tenets which will determine the Army's ability to succeed on the intense battlefield of the future:

initiative
agility
depth
synchronization.

These four tenets, like the previous dynamics of combat power, are the underlying concepts which we believe that a model and measurement method for unit performance must be able to capture and explicate.

"Initiative means setting or changing the terms of battle by action" (FM 100-5, page 15). It may be the most important of the four tenets because it captures the basic attitude which Army leaders at all levels must take to succeed--they must be able to assess the situation and take appropriate prudent risks to achieve their objectives. This means finding a balance between centralizing and decentralizing control that enables forces at all levels to react to conditions and circumstances on the battlefield in the context of overall operational and strategic plans. But how can initiative be measured in the case of combined-arms task force operations? No check list is adequate, nor can a checklist or task-list approach capture the dynamics of the situation in which initiative must be measured.

Agility is the ability to act faster than the enemy is able to react, to keep him off-balance. It requires leaders at all levels who are able to read and react instantly to changes in the battle, who can process intelligence information as it is received, sort out what is important, and act decisively. While individual skills which are required to act with agility can be delineated and presumably measured, the real meaning of the concept must be measured in the context of the battle, as an answer to the question, "Has this unit (leader) acted with agility in this situation?" Only by looking at the whole flow of the battle can agility be measured.

Depth refers to the non-linear nature of the AirLand Battle doctrine, and will be reflected in plans and operations which seek to attack the enemy, not only at the front line, but deep within his rear area. The purpose of the depth strategy is to disrupt enemy command and control, logistics, and movement--to keep him off-balance and thus to create opportunities to exploit with own forces. Achieving depth requires commanders to assess and respond to opportunities as they present themselves, and to use these opportunities to create momentum for U.S. forces. It can be evaluated only by examining the total picture of the engagement and ascertaining whether or not the commander has taken the initiative to identify and attack opportunities that are presented by the particular situation.

Finally, "synchronization is the arrangement of battlefield activities in time, space, and purpose to produce maximum relative combat power at the decisive point" (FM 100-5, page 17). Achieving synchronization requires the commander to be able to visualize the battlefield in all three physical dimensions plus a time dimension, so he can plan and marshal his forces appropriately. It requires both clear planning and the ability to communicate these plans to subordinates who will be forced to act with some considerable discretion within that plan in order to take advantage of

the fast-moving changes on the battlefield. The commander (and subordinates) must be able to recognize and react to the flow of events on the battlefield to achieve synchronization. Measuring the amount of synchronization achieved in an engagement or training exercise requires the application of expert judgment; it is one of the major evaluations performed by the observers at the NTC (Word, 1987).

The principles of war. Principles of war were initially formulated by British Major General J.F.C. Fuller after World War I, and they appear now in Army doctrine with only minor changes from General Fullers's original formulation. These principles are major concepts for the planning and conduct of combat, and are taught to every Army leader. They form the basis of most training that prepares unit leaders, and their application should be recognizable in the actions of units in a combat engagement.

The Army currently recognizes nine Principles of War:

- *Objective*: Direct every military operation towards a clearly defined, decisive, and attainable objective.
- *Offensive*: Seize, retain, and exploit the initiative.
- *Mass*: Concentrate combat power at the decisive place and time.
- *Economy of Force*: Allocate minimum essential combat power to secondary efforts.
- *Maneuver*: Place the enemy in a position of disadvantage through the flexible application of combat power.
- *Unity of Command*: For every objective, ensure unity of effort under one responsible commander.
- *Security*: Never permit the enemy to acquire an unexpected advantage.
- *Surprise*: Strike the enemy at a time and/or place and in a manner for which he is unprepared.
- *Simplicity*: Prepare clear, uncomplicated plans and clear, concise orders to ensure thorough understanding.

As with the concepts discussed above, these principles are not easily measured or evaluated. They all can be characterized as reflecting the patterns and flow of the engagement, and as such will require an approach to measurement which is able to capture the dynamics of the modern battlefield. Measurement techniques based upon sets of tasks to be performed cannot capture these dynamics adequately, although they can be used to evaluate the underlying skills which are required to successfully execute the engagement.

We believe that subject-matter experts watching and evaluating battles and training exercises are able to evaluate these concepts. Review of after-action reports from exercises and the report of COL Word suggest

that the observers of the NTC engagements couch their comments in just such terms. The doctrinal concepts such as those listed above provide the key organizing perspective for evaluation of unit performance. A model for predicting unit performance needs, therefore, to be based upon a similar approach, one which is able to incorporate the doctrinal concepts that Army leaders are taught and which underlie all tactical and materiel developments. The methodology selected must be able to deal with these concepts as the patterns of activity on the four-dimensional battlefield. This review suggests that an approach based upon natural analogy can serve this requirement, and that models of unit performance based upon neural networks can produce the measures and predictions of unit performance being sought.

NEURAL-NETWORK MODELS

Overview of Natural-Analogy/Neural-Network Models

In recent years, advances in computer technology, artificial intelligence, and in the understanding of mental processing have led to the development of a new approach to solving problems of the type being considered here. Known variously as parallel distributed processing (PDP) or neural networks, these models are based on an analogy to the functioning of the human brain and the central nervous system in which complex processes are occurring almost simultaneously, linked by a complex series of connections (in the case of the brain, these connections are composed of biochemical and electrical activity between the nodes). One reviewer has used this analogy to define a neural network as "a computing system made up of a number of simple, highly interconnected processing elements, which processes information by its dynamic state response to external inputs" (Caudill, 1987). The comparison is drawn to a serial computer which processes one instruction at a time; the neural network or PDP concept does not require this sequential set of processing steps but posits that its nodes respond in parallel to a series of inputs presented to it, and that the the result is not stored in some specific memory location but is represented by the state of the system at some equilibrium point. In other words, underlying the concept of a neural-network model is the notion that the pattern of elements that comprise the network is the measure of the status of that network at a point in time.

Neural-network models have found wide-ranging applications in recent years. Most applications have involved some form of pattern-recognition activity, such as inspection functions in an assembly line, review of credit and life insurance applications, computer vision, and voice recognition software. Unlike the large-scale digital and serial processing applied to many large database problems, neural networks rely on many simpler processors, linked together, and modified so that a particular input will produce a desire outcome. The network relates an input pattern to an output in a statistical rather than an exact manner. The key element about these neural-network models is that they can "learn"--that is, they can be "trained" to recognize patterns of input data and to make correct inferences from these patterns.

How a Neural Network Works

A neural-network model consists of a set of *units*, a pattern of *connections* between units, and a rule for *propagating* activation levels through the connections. Units may take on either continuous (typically ranging between 0 and 1) or discrete (usually 0 or 1) activation levels. A common class of activation functions is the *semilinear activation functions*. In a semilinear activation model, activation propagates through the network as follows (Figure 1). The net input of a unit is computed as a function (usually linear) of the connection strengths and activation levels of the units feeding into it, plus a bias term (representing a base level of activation):

$$input_i = \sum_j w_{ij}x_j + bias_i \qquad (1)$$

In this equation, the sum j runs over all units that have direct connections feeding into unit i, and x_j stands for the activation level of unit j. The term $bias_i$ is the "base input," or the level of $input_i$ when it is receiving zero input from the units connected to it.

Next, the unit's new activation level is computed as some function of its net input:

$$x_i = f(input_i) \; . \qquad (2)$$

Some commonly used activation functions include the following (McClelland and Rumelhart, 1988):

Linear: $f(z) = z;$

Linear threshold: $f(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{otherwise;} \end{cases}$

Stochastic: $f(z) = \begin{cases} 1 & \text{with probability } [1 + exp(-z)]^{-1} \\ 0 & \text{otherwise;} \end{cases}$

Logistic: $f(z) = [1 + exp(-z)]^{-1} \; .$

A neural network's knowledge about a task resides in the network topology (i.e., the pattern and direction of links between units), and the connection strengths and biases determining the net input to a unit from its neighbors. Networks may use *local* representations, in which each unit stands for an individual symbol or concept, or *distributed* representations, in which symbols or concepts emerge as epiphenomena of *patterns* of activation on individual, subsymbolic units. A vigorous dialogue is taking place within the connectionist community about the importance of distributed representations. On philosophical grounds, PDP (parallel distributed processing) purists object to local representations. (C.f., the well-known "grandmother neuron" objection to local representations: in the human brain, no single neuron represents the concept of "grandmother.") Yet a number of models using local representations have produced interesting results and contributed to our understanding of connectionist modeling (e.g., Rumelhart and Zipser, 1986). And it should be noted that the "microfeatures" that form the basic units of distributed representations are often symbolic entities, albeit at a level lower than the one of primary interest. Therefore, the debate can be reduced to a discussion of the appropriate level of representation for the specific requirements of a given problem. The proper level depends on computational constraints, robustness requirements, and the availability of a suitable microfeature representation for a problem.

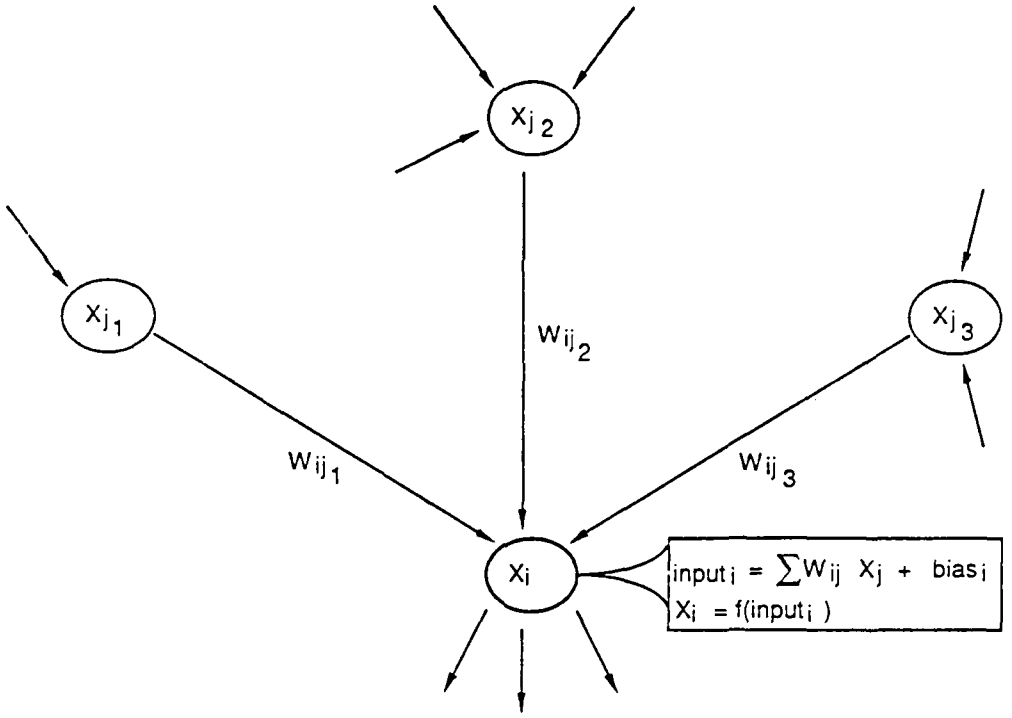

Figure 1. Propagation in a neural network.

The network's knowledge--its topology and connection strengths--can be determined in several ways. The network can be hard-coded for particular problems (e.g., Hopfield, 1982; McClelland and Elman, 1986); a network with randomly assigned connections can be trained using some learning algorithm (e.g., Hinton and Sejnowski, 1986; Rumelhart, Hinton, and Williams, 1986); or some combination of initial structure and learning can be used (e.g., Laskey, forthcoming).

Learning in Neural Networks

Training from a "blank slate" (with only network structure as a constraint) tends to be well suited for low-level problems (such as sonar classification--Gorman and Sejnowski, 1988a,b) in which *a priori* decomposition of the input into a symbolic (or even microfeature) representation may be very difficult. "Hard coding" and mixed methods seem most suitable for higher-level problems for which a symbolic or microfeature structure may already be available, and in which a number of layers (or levels of complexity) need to be represented. It is not clear, for example, whether training a complex multilevel structure such as a semantic network with inheritance (Shastri, 1989) from a blank slate is feasible. Our only model of such learning is humans. Humans take a long time to acquire such structures, require structured training, and receive feedback (via language) at a number of levels, including explicit instruction about the symbolic structure of the inheritance hierarchy.

Learning methods for neural networks can be broadly classified into three categories (DARPA, 1988): unsupervised, self-supervised, and supervised. In unsupervised learning, the system discovers natural categories in unlabeled input data. In self-supervised learning, the system monitors performance internally with no external teacher, and adjusts its model for better performance. Self-supervised learning is appropriate for problems in which the system can generate an internal error signal (such as training a robot arm to reach and grasp an object in its visual field; the system can tell from the visual signal how close to the object the arm is). In supervised learning, an external "teacher" gives the system feedback about how close its response was to some "target" response.

Some examples of neural-network learning mechanisms are described below. Competitive learning (Rumelhart and Zipser, 1986) and ART (Carpenter and Grossberg, 1987) are unsupervised learners that find natural clusters in input data. Darwin III uses self-supervision to follow and touch a moving target with a robot arm (Edelman, 1987). The Boltzmann machine (Hinton and Sejnowski, 1986), the reduced Couloumb energy classifier (Reilly, Cooper, and Elbaum, 1982), and back-propagation learning (Rumelhart, Hinton, and Williams, 1986) are examples of supervised learning systems that can be used for pattern classification. Back propagation is emerging as perhaps the most popular learning technique for problems in which labeled training data exist. We focus in some detail on this model because it is the technique applied in the models of unit performance developed during this research.

Back-Propagation Learning

Back-propagation is a generalization of the delta rule for learning in perceptrons, the original connectionist models (Rosenblatt, 1958). While the delta rule applies only to two-layer systems, back propagation can be applied to networks with intermediate layers of "hidden" units, which provide extra dimensions needed to encode complex patterns.

Back propagation works with any differentiable semilinear activation function (i.e., one in which a unit's activation is a differentiable function of a linear net input). The back-propagation algorithm supplied with McClelland and Rumelhart (1988) uses the logistic-activation function described above.

The basic back-propagation algorithm works for a particular network topology called a layered feedforward network (but it can be generalized to recurrent networks). A layered feedforward network consists of an input layer, zero or more layers of hidden units, and an output layer (Figure 2). Connections run from each layer to the next; no connections may loop back to previous layers. The input layer is set from outside the system, and consists of a representation of the information in the signal to be classified (and, optionally, additional information to be used for classification). The propagation algorithm sweeps through the layers, from input through the hidden layers to the output layer. It uses the activation values in one layer to compute the activation values in the next. The final values computed are the activations in the output layer, which represent the system's response to the classification task.

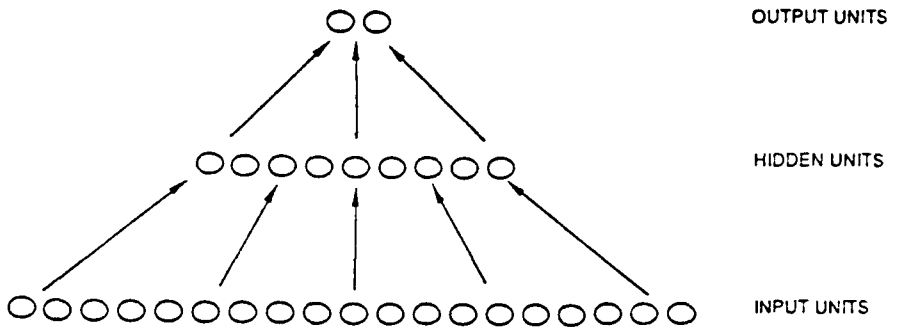

Figure 2. Schematic diagram of a layered feedforward network.

Back propagation gets its name because it propagates the error in the output back through the layers in reverse sequence (i.e., from output to input). The algorithm implements gradient descent in weight space--that is, it changes weights on each iteration by an amount proportional to the gradient of an error function that measures how well the neural network produces the target outputs for given inputs. The error function E is the squared differences, summed over the patterns, between the output of the network and the "target" values for the patterns. The error for a given pattern is given by:

$$E_p = \sum_i (t_{pi} - x_{pi})^2 . \tag{3}$$

In (3), the summation ranges over the output units i. The value t_{pi} is the target value for output unit i on pattern p, and x_{pi} is the activation value of output unit i when presented with pattern p. The total error to be minimized is given by summing (3) over all patterns presented to the system.

After a pattern is presented to the system, the system computes the gradient of the error function for that pattern, and changes the weight in the direction of the gradient:

$$\Delta w_{ij} = - k \frac{\partial E_p}{\partial w_{ij}} . \tag{4}$$

In other words, the weight is changed in a direction that decreases the error E_p. The weight change can be reexpressed (McClelland and Rumelhart, 1988) as:

$$\Delta w_{ij} = \epsilon \delta_{pi} x_{pj} . \tag{5}$$

The term ϵ is an adjustable learning rate; the term x_{pj} represents the activation level of the sending unit j when pattern p is presented; and the term δ_{pi}, the proportional weight adjustment, represents the contribution of a change in the weight w_{ij} on the total error. This term δ_{ij} is computed recursively for each layer in the network using the values for the previously computed layer.

If unit i is a target unit, then:

$$\delta_{pi} = (t_{pi} - x_{pi}) f'(input_i) \quad . \tag{6}$$

Combining (5) and (6), we see that the change in weight w_{ij} after pattern p is presented is proportional to the product of three factors. The magnitude of the weight change increases as each of these moves away from zero. The first factor is the activation level x_{pj} of the sending unit j: an active sending unit increases the magnitude of the weight change. The second factor is the difference between the output unit's activation x_{pi} and the target value t_{pi}: an output that differs greatly from the target increases the magnitude of the weight change. The third factor is the derivative of the activation function with respect to the net input of the unit: a steeply sloping activation function increases the magnitude of the weight change.

When unit i is an internal hidden unit, then:

$$\delta_{pi} = \left(\sum_k \delta_{pk} w_{ki} \right) f'(input_i) \quad . \tag{7}$$

Equation (7) differs from (6) only in the term for the difference between target and output: hidden units have no externally specified target. To obtain an analog for this term, the proportional weight adjustments δ_{pk} are propagated backward from output units through the layers of hidden units. The weight adjustments δ_{pk} propagate to δ_{pi} in the previous layer in proportion to the weights w_{ki}.

An equation similar to (5) can be applied to change the unit thresholds as well as the weights.

The following are the steps in back-propagation learning:

1. Initialize the weights and thresholds in the network. (Often, the network is initialized with randomly assigned weights and thresholds.)

2. Present the network with a pattern. Propagate the pattern forward through the network to the output units.

3. Propagate the error backward through the network, using equations (4), (5), and (6) to compute all weight changes Δw_{ij}.

4. Repeat steps 2 and 3 over a large number of patterns. Usually a training set is presented again and again to the network until the improvement in the total error on each step becomes small.

Building a Neural-Network Model

An experienced Army officer's ability to evaluate the performance of a unit involves the ability to organize raw input data (positions and movements of individual units or vehicles) into patterns, and to relate these

patterns of inputs to success or failure of the unit. An officer's verbal articulation of this process is in terms of high-level linguistic categories--such as the Principles of War or components of Airland Battle Doctrine. Each linguistic descriptor (e.g., "making good use of maneuver;" "poor synchronization") represents a pattern formed from the interaction of a large number of input stimuli.

Recognition of complex patterns like these is one of the promising application areas for neural-network technology. The steps in constructing a neural network for unit performance prediction are as follows:

1. Define the problem the system will solve. What will be the system's inputs? At what point in the progress of the battle, exercise, or simulation will the system make its prediction of performance? What measure or measures of performance will the system be expected to predict?

2. Design a neural-network model and training procedure. Select a representation for network inputs and outputs. Decide how raw input is to be preprocessed before being fed to the network. Define the network topology and any "hard-wired" constraints on connections or weights, and select a form for the activation functions and propagation rules. Select a training procedure. Define what type of feedback at what levels is to be provided to the network.

3. Train the network. Gather a set of training instances representative of the problems the network is expected to solve. Present these instances repeatedly to the network until it learns to classify the instances in the training set.

4. Use the network to classify new instances. Present new instances as they are obtained. The network will classify these instances using the model it has learned from the training instances. If desired, training can continue on the new instances, or the network can be "fixed" after the training sample is complete.

5. Evaluate network performance. Evaluation should be based on how well the system performs on test items it has not seen before (evaluating on the basis of the training sample can give overly optimistic conclusions).

The application of these steps to build a neural-network model for predicting unit performance is discussed in the next section.

PROTOTYPE MODELS OF UNIT PERFORMANCE

Approach

The goal of the Phase I research was to investigate the potential utility of natural analogy or neural-network models for measuring and evaluating unit performance, especially for application to units undergoing training at the National Training Center. Our basic research strategy was therefore as follows.

Step 1: A model of unit performance must operate from information about units that can realistically be made available to the model. The first step, therefore, was to examine the kinds of data that might be available for operating, and especially for training, a neural-network model of unit performance. Two important sources of information about unit performance have been described above: data from training exercises (NTC or JRTC), and data from SIMNET training simulations. Examination of the types of data likely to be available from each, and analysis of the kinds of processing that would be required to make them suitable for neural-network simulations, was a major focus of this preliminary work.

Step 2: The next step was conceptual design of a neural-network approach to unit performance measurement, and implementation of a small-scale prototype model. The prototype was designed to operate from the kinds of data likely to be available to such a model, either from field or SIMNET training exercises. For Phase I, very simple models were developed and tested. Data from these models came from two sources: subjectively assessed data of inputs similar to those that could be obtained from available data and results from plays of a commercial wargame. As described below, a simple neural-network model can make useful measurements and predictions of unit performance.

Step 3: The final step was to formulate a research plan for extending this research to a more realistic Army setting and more complex kinds of data. Such neural-network models must necessarily have more complex structure than the ones built for Phase I of this project. We hypothesize that a combination of learning from data and informed hard-coding of structure will be necessary to build a successful neural-network model of Army unit performance. The data requirements will be more extensive than for the small-scale Phase I models; much larger training sets will be required.

National Training Center Exercise Results

As stated above, the first step in the analysis was to identify and develop suitable sources of unit performance data for the models. The first source identified and tested was data from the National Training Center. The General-purpose NTC Analysis of Training Tool (GNATT) is a computer program written by ARI-POM which enables selected data from the INGRES Mission Databases to be displayed on MS-DOS computers having 16-color (EGA) capabilities. GNATT is based on the assumption that visual representation of data is a powerful aid to understanding. The user can view the training-exercise movement and engagement activity sequentially,

change the viewing scale, display various battlefield graphics, and select units and weapon types for color coding.

GNATT is menu-driven, supports a Microsoft mouse, and was written for use by non-programmers. On-line, context-dependent help is available throughout the program. Input data for six engagements were prepared on the ARI-POM VAX computer using existing INGRES programs, then sent to DSC in machine-readable form for analysis. The six engagements represented a wide range of attack and defend missions, and provided the opportunity to review the usefulness and adequacy of the GNATT and NTC data for the analyses required.

The GNATT data were reviewed by carefully watching the plays of the engagements on a full-color EGA screen as the software processed the data. Alternative scaling and color coding of units and equipment were tried to maximize the amount of information that could be retrieved. Working with an initial list of the doctrinal concepts thought to be important, project staff attempted to ascertain whether a human observer would be able to reach conclusions from the screen presentations to evaluate the degree to which the conceptual requirements contained in the doctrinal principles had been achieved. The human observer was, in this case, acting as a proxy for a complete neural-network model which would be able to recognize the movements and locations of forces from the data being processed by the GNATT model.

For example, we assumed that the principle of mass could be measured by gauging the concentration of forces on the display, but we learned as we proceeded that the instrumented data was often incomplete or inaccurate. The data files from which GNATT is driven capture the position of each instrumented vehicle or person every five minutes through the course of the engagement, yet in many cases vehicles appeared and disappeared, presumably because the instrumentation was not able to receive position locations from vehicles because of terrain features or mechanical errors. Vehicles which had been "killed" appeared to be reactivated in subsequent time periods. As a result, we were not confident that what we were watching was an accurate portrayal of what actually took place during the engagement.

Fire-exchange rates were also thought to be important observables for understanding the pattern and flow of the battle required for developing and instructing a neural-network model, but the instrumented data were quite obviously incomplete on this dimension. Vehicles would be "killed" with no record on the data file of fire exchanges.

Finally, a database of hundreds of cases would be required to provide a training set and a holdout sample for testing the network. To reduce the effect of confounding factors, an initial model would ideally require training cases consisting of different units in the same scenario on the same terrain, but it was not possible to assemble a data set of this size and consistency from the data available. For these reasons it was necessary to evaluate alternative data-gathering approaches to the problem.

Subjective Assessment from Observable Measures of Unit Performance

We hypothesized that certain observations of a battle in progress can be used by an "expert" observer to predict the outcome of the battle. One measure for each of the nine Principles of War was developed. Each of these measures was designed to be calculable from the GNATT data. (As noted above, the data-processing requirements for actually calculating the measures were prohibitive for Phase I.) The following measures were used (precise definitions are given in Table 1):

- *Objective* - direction of attack
- *Offensive* - engagement ratio
- *Mass* - force ratio
- *Economy of Force* - sector ratio
- *Maneuver* - speed
- *Unity of Command* - span of control
- *Security* - air defense protection
- *Surprise* - readiness of defense
- *Simplicity* - scheme of maneuver

Table 1

Notional Test Definitions and Data

MEASURES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. OBJECTIVE - DIRECTION OF ATTACK 180° - DEFLECTION DEGREES FROM OBJECTIVE	9	8	7	6	5	4	3	2	1	0	0	1	2	3	4	5	6	7	8	9
2. OFFENSIVE - WGTD BY WEAPON SYSTEM RATIO OF VEHICLE ENGAGED (% DEF/% ATT)	6	7	1	0	9	4	3	5	8	2	9	8	7	0	2	1	5	4	3	1
3. MASS - WGTD BY WEAPON SYSTEM FORCE RATIO (ATT/DEF)	7	8	5	6	2	0	1	6	7	2	5	5	3	1	5	8	7	6	9	2
4. ECONOMY OF FORCE - SECTOR RATIO (MAIN/SECONDARY ATT)	9	7	0	5	2	8	1	6	0	4	4	3	8	7	5	2	3	5	4	8
5. MANEUVER - SPEED (AVG KM/HR)	9	5	9	8	7	2	1	8	2	4	3	9	6	3	3	7	4	5	0	1
6. UNITY OF COMMAND - SPAN OF CONTROL (% UNIT WITHIN DIST)	8	2	5	9	4	7	5	6	0	2	1	5	9	6	7	0	2	9	8	1
7. SECURITY - ADA COVERAGE (% FORCE COVERED)	5	7	7	6	9	8	0	1	6	5	3	3	4	6	8	9	1	1	4	8
8. SURPRISE - DEFENSE READINESS (% OF DEF MOVING)	4	3	8	9	8	7	4	2	0	2	4	2	5	6	7	7	2	1	0	3
9. SIMPLICITY - SCHEME OF MANEUVER (# GROUPS/DIRECTION)	8	9	9	2	5	6	2	6	5	3	4	2	5	4	3	7	8	7	8	6
OUTCOME	S	S	F	F	S	F	F	S	S	F	S	S	S	F	F	S	S	F	F	F

A set of "notional" data for each measure was defined. Twenty cases with randomly assigned values were created and an outcome was subjectively assessed (success or failure) for each case based on best judgment. Table 1 shows the "notional" data and subjectively assessed outcomes for the twenty cases.

A neural-network model was built using the nine observables as input and the best judgment result as output. The network was trained using back-propagation learning. With this set of data, the neural net was able to learn to discriminate successes from failures; that is, the neural net learned to replicate its training set exactly. No hidden layers were required for the network to learn successfully.

This is one test of the neural network, but a more rigorous test is its ability to correctly classify examples it has not seen before. Therefore, the network was trained on nineteen of the twenty cases, and its ability to predict the case that was not in its training set was tested. Each of the twenty cases was omitted in turn, allowing observation of the network's ability to predict each of the twenty cases from the other nineteen.

In the holdout test, the network was able to classify sixteen of twenty cases successfully. This is a 80% success rate. Thus, this simple neural network was able to duplicate expert judgment on cases other than its training set 80% of the time.

Neural-Network Model of Commercial Wargame Outcomes

The war game. The subjective data set used above might have been contaminated by subjective biases and therefore a more objective test of the neural-network approach was required. A commercial wargame called TOBRUK (1975) was selected to be a second source of data. This manual wargame had the advantages of being simple to learn and to modify for our purposes, and, in addition, was easy to observe and to record data. Further, the desert terrain of the game board closely resembled the terrain found at the NTC.

TOBRUK recreates, in a realistic fashion, the tactical-level combat problems encountered by the British Commonwealth, German and Italian forces confronting each other in the North African desert in May and June of 1942. TOBRUK is played in turns, each representing 30 seconds of elapsed time. Armor, infantry and artillery employments are governed by a set of rules presented for each scenario which clearly identify allowed and prohibited actions by each player. One or more players is designated as being the British Commonwealth side and one or more players is designated as being the German-Italian or Axis side. In general, at the beginning of a scenario, each side holds or enters into one portion of the board and must maneuver its units and engage in combat with units from the other side until time runs out or specified victory conditions are met. These victory conditions are different for each scenario and are intended to "balance out" the scenarios, thus making results as much dependent on player skill as possible.

The TOBRUK mapboard is an hexagonal grid representing the flat, featureless desert terrain where the actual campaign occurred. The hexagonal grid system is used to regulate movement and combat. The shaded hexagons (called "hexes") are used to indicate area boundaries for unit placement in the various scenarios. A grid-coordinate system, printed around the edge of the mapboard, consists of letters running north-south and numbers run-

ning northeast-southwest. Hex locations are identified by cross-referencing a letter hex row with a number hex row.

Central to play are the chance dice and probability charts and tables. The most important of these are the Hit Probability Table and the Damage Table. When a player is ready to engage one of the other player's vehicles with one of his own, he rolls the dice and looks on the first table to determine whether he has hit the enemy vehicle. If so, he rolls the dice again to determine the extent of the damage. Vehicles may be rendered immobile, incapable of returning fire or totally destroyed.

At each turn, the players can maneuver or fire their vehicles, and then estimate the results using the probability tables calculated from actual World War II performance data and provided with the game. For this data collection, two players (who alternated between attacking and defending) and an observer (who collected data on moves and outcomes for input into the neural-net model) were used. Vehicles were limited to tanks and infantry fighting vehicles for simplicity.

Measures of unit performance. A few games of TOBRUK were played initially to acquaint the players and the observer with the game and to identify possible measures of performance that could be used to predict outcomes in a neural-net model. Fourteen measures were identified:

1.	Mass	Attacker to defender vehicle ratio at end of turn
2.	Mix	Number of attacker tanks/total attacker vehicles at end of turn
3.	Firepower	Number of attacker vehicles fired in turn/ number of defender vehicles fired in turn
4.	Mobility	Total hexes moved by attacker/total attacker vehicles at beginning of turn
5.	Objective	Absolute value of difference in average attacker ing and direction to objective in radians
6.	Dispersion	Average hexes between attacker vehicles
7.	Synchronization	Variance in heading
8.	Frontages	Standard deviation of widths of attacker/standard deviation of widths of defender
9.	Security-Attacker	Percent of attacker with flanks protected at beginning of turn

10.	Security-Defender	Percent of defender with flanks protected at beginning of turn
11.	Defense Readiness	Percent of defender vehicles not moved in turn
12.	Key terrain	Number of prepared weapons emplacements
13.	Obstacles	Number of hexes blocked by obstacles
14.	Obscuration	Number of hexes obscured by smoke

Note that these measures were not the same as the ones used for subjective assessment. This is because the measures reported here had to be calculable from the TOBRUK board positions. However, all of these variables are comparable to the types of variables that are available from the NTC or SIMNET systems.

Thirty games of TOBRUK were played and recorded. The games used a single scenario in which the Allies defended a critical asset (assumed to be a fixed-communications post) and the Axis forces attacked to seize the position. Each of the above measures was calculated and recorded on the third turn after the first engagement of the game. The game was played by four different individuals, with participants playing different roles. Each game ended as either a success or a failure for the attacker. The results of the 30 games, and the assessed values of the 14 variables, are displayed in Table 2.

Results using all variables. A neural-network model was fitted to the data from the plays of the game, with the fourteen measures as input and win/loss of the attacker as the target output. As for the subjectively assessed data, a neural-network model with no hidden nodes correctly reproduced the training set with 100% accuracy.

The holdout sample test, however, yielded less satisfactory results. Each case in turn was left out of the training sample. The net trained on the other 29 cases was then used to predict the holdout case. Unfortunately, the network identified the left-out case only 13 out of 30 times, for a success rate of only 43% on cases other than its training set.

There were, we hypothesized, several reasons for this inability to predict. First, the subjectively assessed outcomes were *modal* outcomes. In other words, the outcome for a set of measures was assessed to be a victory or failure if that outcome was judged *most likely*, given those inputs. But clearly there is uncertainty associated with performance given these inputs. Indeed, for many games, TOBRUK players felt that victory or defeat was not determined until the final moves of the game. So the TOBRUK data were inherently noisier than the subjective data.

Second, the network was deliberately given "hard" cases. After the first few rounds of the game, it became clear that one could determine the output of the game by giving one side sufficient mass. Initial conditions after the first 3 or 4 games were set in a range that allowed more of the

Table 2

TOBRUK Input Data

Case/Measure	Mass	Mix	F'Power	Mobil	Objec.	Disper	Synch	Front	Sec-A	Sec-D	Def Red	Terr	Obsta	Obscur	OUTCOME
1	1.3	1.0	0.7	2.3	0.1	5.0	0.8	0.5	0.8	0.9	0.8	0.4	34.0	0.0	F
2	2.0	1.0	2.0	0.0	0.1	10.8	0.8	2.9	1.0	0.8	1.0	1.0	0.0	0.0	F
3	3.6	0.5	1.0	1.6	0.2	7.1	0.6	2.2	0.9	1.0	1.0	1.0	6.0	0.0	F
4	3.4	0.4	3.3	0.7	0.7	3.0	1.0	0.2	1.0	1.0	1.0	0.5	6.0	0.0	S
5	3.8	0.5	1.8	1.6	0.4	11.4	0.6	1.8	1.0	0.6	0.8	1.0	6.0	0.0	S
6	2.8	0.4	1.0	1.6	0.2	9.8	0.8	1.0	1.0	1.0	1.0	1.0	8.0	1.0	S
7	3.3	0.4	3.0	1.2	0.7	6.5	0.8	0.7	1.0	1.0	1.0	1.0	8.0	2.0	S
8	2.3	0.3	1.5	1.2	0.4	4.8	0.8	0.8	0.4	0.8	0.8	0.7	8.0	3.0	F
9	1.8	0.4	1.0	1.2	0.1	12.9	0.8	1.4	0.6	1.0	1.0	1.0	8.0	2.0	F
10	5.0	0.3	1.0	2.1	0.2	5.6	1.0	1.7	1.0	1.0	0.0	0.0	8.0	2.0	S
11	3.3	0.3	1.5	1.4	0.2	3.5	0.8	0.7	1.0	0.7	1.0	1.0	8.0	0.0	S
12	2.2	0.3	2.0	1.7	0.5	5.0	1.0	0.7	1.0	1.0	1.0	1.0	3.0	4.0	F
13	3.8	0.7	1.0	2.1	0.1	6.8	1.0	1.4	0.7	0.3	1.0	1.0	8.0	0.0	S
14	2.5	0.7	2.0	1.5	0.2	7.1	0.8	1.1	0.9	0.8	1.0	0.7	5.0	5.0	F
15	3.5	7.0	0.0	2.1	0.0	7.1	0.6	1.6	1.0	1.0	1.0	1.0	12.0	1.0	S
16	1.8	0.6	0.5	1.3	0.4	8.9	0.6	0.9	1.0	1.0	1.0	0.7	6.0	1.0	F
17	3.3	0.9	1.3	1.5	0.2	10.0	0.8	0.7	0.9	0.3	0.7	0.0	20.0	0.0	S
18	2.5	0.7	0.4	1.9	0.2	9.1	0.8	1.3	0.9	1.0	1.0	0.8	12.0	1.0	F
19	3.3	0.8	0.0	2.5	0.0	8.4	0.8	0.9	1.0	1.0	1.0	1.0	15.0	3.0	F
20	6.3	0.1	2.0	1.9	0.0	2.3	1.0	0.4	1.0	1.0	1.0	1.0	5.0	0.0	S
21	5.0	0.4	1.0	1.8	0.0	7.7	1.0	0.7	0.8	1.0	1.0	1.0	5.0	0.0	S
22	4.7	0.3	1.5	1.6	0.1	4.2	0.8	0.7	1.0	1.0	1.0	1.0	9.0	1.0	F
23	3.3	0.5	0.0	1.9	0.2	8.6	0.8	1.2	0.9	1.0	1.0	0.7	2.0	2.0	S
24	1.8	0.6	1.0	0.9	0.0	5.3	1.0	0.7	0.9	0.8	1.0	1.0	18.0	1.0	F
25	2.3	0.6	3.0	1.6	0.5	4.7	0.8	0.5	1.0	1.0	1.0	1.0	23.0	2.0	F
26	3.0	0.1	2.0	0.9	0.1	8.7	1.0	0.9	0.8	0.7	1.0	1.0	19.0	1.0	S
27	3.3	0.2	1.0	0.8	0.1	1.8	1.0	0.4	1.0	1.0	1.0	1.0	4.0	2.0	S
28	3.3	0.7	4.0	1.1	0.1	8.4	0.8	0.9	0.9	0.8	1.0	1.0	11.0	2.0	S
29	2.0	0.7	1.3	1.2	0.2	8.8	1.0	0.9	1.0	1.0	1.0	1.0	7.0	2.0	S
30	2.8	0.2	0.0	1.8	0.1	4.9	1.0	0.6	0.9	1.0	1.0	1.0	13.0	2.0	F

other variables to determine outcomes. Still, a good neural-network model should be expected to be able to perform better than chance. But recall that the network was given only 30 training cases, with 14 input nodes from which to predict the 30 success/failure values. We hypothesized that this was a case of "overfitting."

Explanations for the weights. Table 3 shows the estimated weights for the neural network using all fourteen variables, and using the entire training set. Because of the small size of the training set and our hypothesis of possible overfitting, the following explanations of the weights should be regarded as tentative and subject to change with further research.

Table 3

Weights in TOBRUK Neural Network (Full Model)

Mass	20.24
Mix	29.81
Firepower	- 1.04
Mobility	- 8.04
Objective	6.60
Dispersion	27.92
Synchronization	16.75
Frontage	-35.87
Security-Attacker	8.02
Security-Defender	-18.41
Defense Readiness	0.95
Key Terrain	-10.60
Obstacles	-23.71
Obscuration	-12.02

MASS - Mass is clearly a dominating factor in a successful attack. Mass correlates directly with firepower (see comments under that variable below), so greater mass increases the probability of destroying the enemy. In addition to increasing the number of times one can fire, mass allows for multiple engagements of the same target by many attackers. Mass also allows for a more varied strategy. With more vehicles, the attacker can gain an advantage by flanking his opponent and establishing multiple fronts. A defender with more mass can prevent flanking of his own vehicles. In addition, a greater mass allows the attacker to absorb more casualties while attempting to achieve his objective. As expected, the neural-net model placed a high weight on the Mass variable, indicating a high correlation between the mass ratio and the chance of victory.

MIX - This weight was also highly positive. This makes sense as the German Pzkw-IIIh tanks are faster, deadlier, and better armored than the Italian M13/40's and, with the same absolute number of tanks, one would expect to do better with more Pzkw-IIIh's than with more M13/40's.

FIREPOWER - This weight was in general small in magnitude, and seemed to fluctuate between positive and negative values depending on the case removed for deriving the weights. This might seem surprising--it would appear that being able to deliver greater firepower should mean having a greater ability to destroy the enemy. But (as noted above under Mass) much of this firepower effect may have been accounted for by the Mass variable. Moreover, the conditions under which measurements were taken may have been poorly suited to measuring the independent effect of mass. This is because one could not fire and move one's tank in the same turn. This caused the attacker to have a mixed strategy in respect to firepower and mobility or speed. The attacker had to choose between moving in as fast as possible to overrun the enemy by sheer mass, and stopping to fire before exposing his flank or taking a damaging hit. Also, due to the fact that the Pzkw-IIIh's

were easily M-killed (mobility killed), many would be forced to stop and fire from a long range with a very small chance of killing the defender as they were helpless to do anything else. The defender, on the other hand, rarely moved and had a chance to fire all of his tanks (unless smoked) every turn. The defender was in a position where it made no sense to move as being in a weapon pit gave a defensive advantage, and there was no cost for the number of vehicles the defender lost in protecting the base. Therefore, the defender fought to the death in preplanned positions, whereas the attacker had to throw all his forces at the defender in the hope he could kill enough of them or keep enough of his forces alive to reach the base.

MOBILITY - The weight for mobility was small in magnitude (in fact, depending on which case was left out, it was often negative). This might lead one to believe that mobility was not important and actually a hindrance in trying to win a battle. The importance of mobility is described above. But in fact, mobility is crucial to battle success. The attacker needs to be close enough to deliver firepower effectively, especially to gain the ability to deliver flanking shots. If the attacker charges in, the defender must react quickly and in full force to prevent the attacker from penetrating. Thus, mobility of the attacker forces the defender to play to his tempo: if the attack is sluggish, the defender can sit back and slowly pick off the attacking vehicles one by one (Crouch and Morley, 1989). A lethargic attack also gives the defense time to make any adjustments and to maneuver.

Why, then, was this not reflected in the neural-network weight? First, we noted above the rule not allowing firing and moving a tank in the same turn. Second, the time at which the snapshot was taken was usually after the initial moves when the attacker was charging the enemy and had yet sustained few casualties. By the time the snapshot is taken, many of the tanks are mobility killed and cannot move if desired (one way to adjust this in the future would be not to count the tanks that cannot move in the Mobility/Speed ratio; instead, only those tanks able to move on that turn would be recorded in this measure). In addition, at this point in the battle, strategic decisions need to be made about whether to stop and fire now that the tanks are within range to do damage to the defender.

Mass, Firepower, and Mobility are tightly coupled aspects of the battle. We suspect that our data set was too small to allow us to sort out their independent and interactive effects. Interactive effects would be accounted for by a neural network with hidden nodes. A network without hidden nodes was able to correctly replicate classification of the training set, but we suspect that a larger data set would reveal interactions that would require a network with hidden nodes.

SYNCHRONIZATION - Players felt that this variable should not make a difference. Players always oriented themselves as to receive the least number of flank shots. Being unsynchronized on the offense may actually have signaled that a piece of the front of the defender had been destroyed, thus allowing the attacker to pivot and protect himself while continuing to expose the enemy's flank. The weight for synchronization, however, was positive (but not as high as mass or mix).

OBJECTIVE - The only reason for not heading directly at the objective would be the engagement of the defender. The weight for this variable was positive but small in magnitude.

FRONTAGE - A wider front was probably advantageous due to the flanking opportunities it presented, however, one must have enough mass so that the front is not too thin. If the front is too spread out, the defender may be able to destroy parts of it and then gang up on what remains. The weight for this variable was large and negative. This is probably due to the fact that most of its effect was captured by the high positive weight for mass. The large negative effect then captured the disadvantage of having a front that was spread too thin for the available mass.

DISPERSION - If troops are not well enough dispersed, it creates an easy advantage for the defender to obtain flank shots and a defensive advantage in not needing to worry about their own flanks. The weight for this variable was large and positive.

OBSCURATION - If used at an inappropriate or inopportune time, the smoke may have little effect upon the outcome of the battle. Smoke can be used effectively. But because it remains for only 2 turns (not including smoke caused by burning tanks), the amount or timing of the use of smoke may not show up during the snapshot turn. The weight for obscuration was moderately negative.

OBSTACLES - As briefly touched on in the discussion of Mobility, there is rarely any need for the defender to leave a weapons pit, only not to be occupying it due to destruction. Weapons pits helped the defender, but once the defender was outnumbered, the weapon pit did little good. One tank, in a weapons pit or not, is not going to do much against 7 tanks. The estimated weight for obstacles was large and negative, possibly reflecting this outnumbering effect.

READINESS - As noted, there was little or no incentive for defending tanks to move. At 100% readiness there was probably an equal chance that there would be victory for either side. Who won was probably dependent on other factors. The estimated weight for readiness was very small in magnitude.

SECURITY (ATTACKER AND DEFENDER) - Security Defender should provide a more accurate clue to the chance of succeeding or failing due to the fact that if a defender was vulnerable, it gave the attacker a much greater chance of killing the defender, whereas the Security Attacker was not as important as the attacker was vulnerable from both the front and the flank. The weight for Security Defender was indeed moderately negative (indicating a negative relationship to success on the attack). Security Attacker had a positive weight, but it was small in magnitude.

Results using only critical variables. The next step was to identify the few variables felt to be the most important to predicting success or failure. This assessment was based on the judgment of the players of the game about which variables were most important for predicting success. Five variables were chosen: Mass, Mix, Firepower, Mobility, and Defense Security.

A neural network was trained using only these five input variables. This time, the model was unable to fit exactly its entire training set: after thousands of time steps, it was still predicting three cases incorrectly (or 10% of the total 30 cases). Adding two hidden nodes to the network allowed it to match one additional case in the training set.

The next test of the neural network was to test it on its prediction of each case after training on the other 29 cases. This time, the neural network with five variables and no hidden nodes did much better than the model that used all the measures: it correctly predicted 22 of the cases it had not seen, for a percentage score of 73%. This confirmed the hypothesis that the model with 14 inputs was overfitted.

The final test of the neural-network model was to compare it with human judgment. The game boards at the time at which measurements were taken for the neural network were reconstructed, and two project staff members were asked to predict the outcome. Both staff members were very familiar with the TOBRUK game--in fact, they had played many of the games in the data set and could reasonably be called experts. Note that in this test they had access to more information than the neural network, in that they had a holistic view of the game board, whereas the neural network had only the five performance measures. Neither the human experts nor the neural network could use information about game dynamics (which we hypothesize to be powerful predictors of success). The two human experts correctly predicted 23 and 20 cases, or 77% and 67%, respectively. Thus, the neural network (which predicted 22 cases correctly) did only slightly worse than the better of our human experts, a quite credible performance.

Table 4 cross-tabulates the frequencies of cases classified correctly and incorrectly by the network and each of the experts. Numbers shown in parentheses are fitted values for the model that assumes independence between network and human judge. Non-independence would arise if some cases were more difficult than others for both network and expert, causing correlation between the network's and the expert's classifications. Both experts agree with the network more than predicted by the independence model. But this result was significant for only one of the human experts ($p=.04$ and .24 for Expert 1 and Expert 2, respectively). Moreover, the two experts agreed with each other no more than predicted by chance ($p=.54$).

Table 4

Classification of Cases: Network and Human Experts

		EXPERT 1	
		Correct	Incorrect
Network	Correct	17 (15)	5 (7)
	Incorrect	3 (5)	5 (3)

Table 4 (Continued)

Classification of Cases: Network and Human Experts

		EXPERT 2	
		Correct	Incorrect
Network	Correct	18 (17)	4 (5)
	Incorrect	5 (6)	3 (2)

Analysis of weights. Table 5 shows the correlation of the weight vector obtained when the network was trained using the entire training set with the weight vector obtained when the network was trained with each of the cases left out. When the full-sample network agreed in its classification with the one-left-out network (whether the classification was correct or incorrect), the weights correlated .999 or above. When the one-left-out classification was incorrect and the full-sample classification was correct, the weights correlated on average .920, and never higher than .992.

Table 5

Correlation of One-Left-Out Weights with Entire Training Set Weights

Case Classified Correctly by Both Networks (Average = 1.000)	
Case 1	1.000
Case 4	1.000
Case 5	1.000
Case 7	1.000
Case 8	1.000
Case 9	1.000
Case 10	1.000
Case 11	1.000
Case 12	1.000
Case 13	1.000
Case 14	.999
Case 16	1.000
Case 17	1.000
Case 18	1.000
Case 20	1.000
Case 21	1.000
Case 24	1.000
Case 25	.999
Case 26	1.000
Case 27	1.000
Case 28	1.000
Case 30	1.000

Table 5 (Continued)

Correlation of One-Left-Out Weights with Entire Training Set Weights

Cases Classified Incorrectly by Both Networks (Average = 1.000)	
Case 3	1.000
Case 22	1.000
Case 29	1.000
Cases Classified Correctly by Full Sample Network, Incorrectly by One-Left-Out Network (Average = .920)	
Case 2	.777
Case 3	.973
Case 15	.904
Case 19	.992
Case 23	.955

Table 6 shows the fitted neural-network weights using the full training set. Note that this network was unable to correctly classify three of the cases in the training set.

Table 6

Weights for 5-Variable Model

Mass	52.90
Mix	-1.05
Firepower	5.38
Mobility	-11.52
Security Defense	5.03

MASS - This was clearly the dominating factor. None of the other weights comes close to it in magnitude.

MIX - This variable went from having a highly positive weight (on the full data set) to having nearly a zero weight in the five-variable model. Setting it to zero did not cause the network to change its classification of any cases, indicating that this variable did not help the network's predictive power. This is surprising, given the superiority of Pskw-IIIh's.

FIREPOWER - The magnitude of this weight was small, but it was consistently positive across cases left out (being negative but very small in two instances). Setting its weight to zero caused the network to misclassify two additional cases. As noted above, with the small data set the effect of Firepower probably could not be estimated independently of Mass and Mobility.

MOBILITY - Unlike the simulation using all the variables, the weight for Mobility was moderately negative (recall that it was positive but very small when all variables were included). Setting its weight to zero caused seven additional cases to be misclassified. The negative effect may have reflected the rule that moving tanks cannot fire.

SECURITY DEFENDER - This weight was positive but small. This is the direction opposite to that expected. But setting it to zero caused three additional cases to be misclassified.

Discussion

When all variables were used to fit the model, the resulting network was unable to predict very well outside its training sample. There are several reasons for this. First there were only thirty trial runs and fourteen variables. This is a limited number of trials to determine the possibly complex interactions between all of these variables.

Second, after about the first 5-7 trials, the players had a rough idea of what would create a decisive victory in terms of mass, mix and strategy. Many "decisive" victories were therefore eliminated and this resulted in many battles that went down to the wire. The fact that chance had a greater importance in these cases rather than strategy or sheer numbers would cause the model difficulty in predicting a winner from the snapshot when even the players were unsure several turns later. If many more trials were run, the neural net would be able to predict the winner or give the percent chance that one side would win in that particular situation (in the same way that meteorologist predict the weather by stating "there is a 40% chance of rain today").

Last is the fact that in each case the players tried to vary as many of the factors as possible without biasing the strategy. By adjusting the initial mass, mix, number of weapons pits, smokes and obstacles and the strategy of attack (which affected frontage, synchronization, dispersal, and objective) all but 5 of the variables in the equation were influenced so as to hopefully induce variance in the data sample. (The five that were not altered were Security Attacker and Defender, Firepower, Mobility, and Readiness). Since each case is unique in its characteristics, the neural net relies on its inclusion in the model to come up with weights that will fit all the data. Theoretically, if enough data points are included, situations will begin to repeat themselves and the neural net will be able to predict with more accuracy.

When the number of variables included in the model was reduced to five, the model's ability to predict cases it had not seen was dramatically improved to rival human performance. The cost, of course, was an inability

to correctly classify all cases in the training sample. Perhaps a chance element was at work on these cases, or perhaps they reflected more complex variable interactions that a simple model could not handle (such as TOBRUK data).

An analysis of the weight magnitudes on both the full model and the reduced model clearly showed the decisive effect of the mass variable. Its weight was consistently large and positive. The contributions of the other variables were secondary, and the direction of their influence seemed to depend on which other variables were included in the model. Future simulations should adjust the initial mass to a range in which the model would be unable to predict the outcome from mass alone.

CONCLUSIONS AND RECOMMENDATIONS

The results of this Phase I work suggest that neural-net methodology holds promise as a low-cost substitute to detailed task analysis for predicting unit performance in training exercises such as those at the NTC. The networks developed and tested are able to do a credible job of recognizing and learning patterns of variables which are associated with success in the actual exercise.

Conclusions From Phase I Work

The neural-network model described earlier, despite its extreme simplicity, predicted as well as human judges who could be considered experts at the prediction task. Yet, because of the constraints of the Phase I task, the distinctive capability of neural-network models--their potential for representing dynamic and configural cues that humans report to be important in classification problems of this kind--could not be tested. Because the network operated on a single "snapshot" of the game board, it was not able to represent the "flow" of the battle. Because it was given only simple measures such as equipment tallies, it was not able to represent patterns of vehicle placement, as our hypothetical general on the hill would use to judge the overall quality of a static fighting position. Note that although our human judges had such configural information available to them, our neural network was able to equal their performance without the configural cues. An interesting research question is the degree to which the more complex multi-layered network would outperform a simple network like the one trained here.

A complete neural-network model of unit performance in a combat engagement such as those simulated at the National Training Center will need to be capable of digesting a mass of data and organizing it into understandable patterns which then can be evaluated. This means, for instance, being able to track the movement of every vehicle and system involved in the engagement (or at a minimum, a consistent sample of vehicles) in order to estimate concepts such as mass, direction, velocity, and synchronization. At least theoretically, the MILES and related data-capture systems being installed at the NTC will provided an automated source of those data. Alternatively, a simulation system such as SIMNET will do the same. The first problem for a neural-net model to address is the processing and aggregation of that mass of data to arrive at the structure of the neural-net model.

Because of resource and data constraints in this limited Phase I research, this data-capture and aggregation problem could not be sufficiently addressed. Sufficient research on pattern-recognition algorithms has been completed, however, to convince us that these problems, while of a large scale, can be solved. The patterns which will need to be assembled from the mass of MILES data are relatively simple geometrics which a model should have little trouble aggregating.

In this Phase I, we began at the point in which these intermediate variables had been assembled into patterns--that is, we assumed that

measures of variables such as mass, firepower, and synchronization had already been constructed from raw MILES-type data. The neural-net models built and tested for the two data sets were intended to test whether a simple model of this type could do an acceptable job of learning to predict the outcome of the engagement. In the case of both the hypothetical data and the board-game data, the results of the experiment suggested that a neural-net model was an appropriate choice. Two pieces of evidence described in this report support this conclusion. First, the model did as well as human experts in predicting engagement outcomes in these two tests. Second, the weights for the variables in the model correspond closely to the relative importance of the concepts of AirLand Battle doctrine examined. These two findings suggest that the neural-net models built and tested possess both construct and face validity.

The substantive results of our modeling efforts are further supported by Brigadier General (Retired) William W. Crouch and Lieutenant Colonel Thomas V. Morley who identified the conditions for a successful attack based on their observations of units at the NTC. These were:

- Use tight, agile formation capable of rapid, controlled employment.
- Have constant maneuver that avoids being slowed by improper overwatch techniques.
- Avoid dissipating the unit's mass through separate engagements.
- Concentrate massed artillery fires to destroy or degrade enemy direct fire systems during the assault phase.
- Be able to commit successive units without gaps that can be exploited by the enemy.
- Get on the objective as rapidly as possible to gain the attacking force's kill advantage.

These conclusions made based upon actual training exercises were very similar to the ones based upon results from the simulated battles of TOBRUK. A more complex neural net will provide new insight into which factors influence the course of battle and to what degree.

Implications For Follow-On Work

One critical aspect of neural-network models is that they learn, from data they are presented, to recognize patterns when offered new data. This means that the initial data set from which the model is trained must be large and rich. Therefore, a critical issue for continuing work is access to such a database. As discussed above, the NTC database appears to be noisy and incomplete and may not be useful for this purpose at the present time. (Steps are being taken to improve the instrumentation and data collection at NTC which will allow reconsideration of this issue in the future.) The SIMNET system is a reasonable substitute, however; it produces a data set which replicates a clean NTC data set and, because it

is a networked computer simulation, the noise and missing-data problems are minimized. Initial conversations with other contractors involved in the design, implementation, and operation of SIMNET suggest that this data set will allow the necessary modeling activity to proceed. In particular, SIMNET is able to faithfully replay the same engagement so issues of variability in the data can be more easily controlled.

This still leaves the issue of a front-end pattern-recognition processor for the neural-net model. The data which the model must preprocess is a large, three-dimensional data set which includes the locations and activities of a myriad of vehicles, weapons, and individual soldiers over some period of time. From these data, the net must be able to learn to recognize visual, spatial, and temporal patterns of movement, massing of forces, velocity of the attack, synchronization of the attack and defense, firing patterns, etc. Building such a front-end processor will be a long and complex task, but we believe that it can be aided considerably by the development of a combination of hard coding of judgmental data and a learning network model. This combination approach is described in more detail below. The most likely combination would be an expert judgment-based module which was able to capture the broad qualitative patterns of unit movement to, and initial stages of, an engagement, coupled with a learning network which then processed those data to identify the modeled precursors of engagement success and failure. Such a model could be taught and then operated in a predictive mode, from a data set such as that available at SIMNET.

Follow-On Research Issues

A number of theoretical and practical issues need to be addressed before the promise of neural networks for unit performance measurement becomes a reality. Some of the modeling decisions that need to be made and some of the technical hurdles that need to be overcome to develop a successful neural-network model of unit performance are discussed below.

Output. The first task is to specify the problem the system must solve: specifically, what question the system is supposed to answer. In Phase I, the problem of predicting, from data on vehicle positions, types and movements, the ultimate success or failure of a unit in an NTC-type training exercise was the focus. The conclusion from this preliminary research is that this task is appropriate and technically feasible for neural-network technology. If this is the problem chosen for a more ambitious neural-network system, several questions remain to be asked, including: (1) Is the system's response supposed to be unidimensional (e.g., success vs. failure) or multidimensional (e.g., using different measures of success or failure)? (2) How should success be measured? (3) Should the system respond with a yes/no answer, or with a graded response indicating the probability of success?

Input representation. A neural network's ability to learn can be greatly influenced by the way in which input data is represented. For example, Gorman and Sejnowski's (1988a) sonar signal recognition system was given input obtained by preprocessing raw sonar signals using a Fourier filtering method. This preprocessing technique was developed using ex-

perimental data about how the human perceptual system processes signals. It is likely that the system's good performance was in part due to this input preprocessing. Many neural-network systems for visual tasks preprocess inputs to achieve location, rotation, and scale invariance. Systems that can tolerate translation or deformation (e.g., Fukushima and Miyake, 1982) require a very large number of processing units. Network inputs must be readily computable from raw inputs (in this case, data such as vehicle location, vehicle type, and temporal marker).

Network structure. Another determinant of network performance is the structure of the network. This includes the pattern of connectivity and the form of the activation function. The more unconstrained the pattern of links, the longer the training time tends to be (because of the large number of parameters to be learned) but the more general the problem the system can eventually solve.

An active area of research in the neural-network community is the development of modular systems. Separate networks can be designed for relatively separable problem components, and "pieced together" to solve more complex problems. For example, one network might be trained to recognize different maneuver patterns; another might be trained to recognize the pattern or timing of an attack. After training, these pieces could be incorporated as components in a larger system that judges overall unit performance.

Different learning algorithms pose different requirements on the structure of a network. The Neocognitron (Fukushima and Miyake, 1982) assumes a particular layered structure and a particular form of propagation of activation between layers. The standard back-propagation technique requires a layered feedforward network (all links go to nodes in the next layer from the sending node; no backward or within-layer nodes are allowed) and a differentiable semi-linear activation function (Rumelhart, Hinton and Williams, 1986). The technique has since been generalized to some kinds of recurrent networks. As described by Hinton and Sejnowski (1986), the Boltzmann machine operates on a completely connected network with a certain probabilistic form for the activation function. The technique applies to arbitrary connectivity patterns, and can be generalized to different probabilistic activation functions (Laskey, in preparation).

Training the network. Important questions to be considered are: (1) Which training technique should be used? (2) What kind of feedback should be provided to the system? (3) What kind and number of training instances should be provided?

The first and second questions interact, because different training mechanisms make different assumptions about what kind of feedback (if any) is provided to the network. In the context of measuring unit performance, it would seem necessary to provide the system with labeled training data. That is, if the system is expected to learn to distinguish between success and failure of a unit, it needs to know whether each input corresponds to a success or failure. It may also be useful to give it other information (e.g., a modular system would need feedback to train each submodule).

Given that a supervised learning technique is appropriate, the question remains as to which technique to use. In Phase I of this project, back propagation was applied. The reason for this choice is the technique's popularity, its suitability for the problem, and the ready availability of software (McClelland and Rumelhart, 1988). This technique is a strong contender for a more ambitious unit performance network to be developed in future research. But there are problems with the back-propagation technique. One problem is that learning can be very slow when there are many layers. This problem is, we believe, fundamental. It must be expected that complex patterns will take a long time to learn, and other learning techniques in fact face the same difficulty. There are possible ways to lessen the severity of the problem, and future research will explore these. Making the system modular and training the pieces separately is perhaps the most promising idea. Of course, its success will depend on how the problem is split up. Researchers are working on modular systems, but we know of no general results as yet on guidelines for how to modularize a problem. Success of training will also be largely dependent, both on the pattern of connectivity imposed on the system and on the input representation. Both these factors are described above.

A second issue is convergence properties of the back-propagation algorithm. The algorithm is a gradient-descent technique, and gradient-descent algorithms can get stuck in local minima or begin to oscillate. When there are no hidden nodes, it is guaranteed to converge to a solution (Rosenblatt, 1958). However, when there are hidden nodes, these results no longer apply. When hidden nodes were added to the networks, oscillations were sometimes encountered (but we do not know whether, if we had waited longer, the system would have stopped oscillating).

A final issue is the number and kind of training instances required to arrive at a solution. This issue is discussed further below.

The training data. The training data given to the system must be representative of the range of problems it is expected to solve. The system must be given enough similar cases that it does not arrive at spurious generalizations. This problem was encountered in the sample network. When 29 training instances were provided for a network with 14 connections to learn, the system performed poorly (correctly classifying less than half the holdout cases). But a system with five connections (using five carefully chosen inputs) performed as well as our human subjects.

In other research, networks have typically been given hundreds to thousands of training instances. Such quantities of training data will be very difficult to obtain from NTC exercises, but may be available from SIMNET runs. It may be possible to construct a simulation that generates large numbers of training instances, build a model based on the simulation, and then use that model structure as a starting point for training a system on more realistic data. If the structure and connectivity of sub-modules built from a simulation can be tested, the amount of training data required from the field might be drastically reduced.

Neural-network learning, as currently applied, is very bottom-up. From a very large number of training instances, the network learns how to classify future instances correctly. This works well for simple problems,

especially when the input representation and network topology can be constructed to exploit the problem structure. But on complex tasks, successful humans learn by combining the bottom-up and top-down modes. That is, learning is most successful when training instances are presented with explicit instruction on useful classification principles. Devising top-down instruction for neural networks is difficult, because a network's internal structure is generally opaque to humans. Analyses of the internal structure of even simple networks has taxed researchers' ingenuity (e.g., Gorman and Sejnowski, 1988b). Modularizing networks is one way of approaching top-down instruction--feedback can be provided both on the "final" output and on the output of the modules providing intermediate answers.

REFERENCES

Carpenter, G.A. and Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54-115.

Caudill, M. (1987). Neural Networks Primer. AI Expert, 2.

Crouch, W.W. and Morley, T.V. (June 1989). Failed Attacks and Flawed Overwatch. *Military Review*, pp. 15-25.

DARPA (1988). *Neural network study*. Fairfax, VA: AFCEA International Press.

Dupuy, T.N. (1987). *Understanding war*. New York, NY: Paragon House Publishers.

Edelman, G.M. (1987). *Neural Darwinism*. NY: Basic Books.

Forsythe, T. (1987). *A Research Concept for Developing and Applying Methods for Measurement and Interpretation of Unit Performance at the National Training Center*. Monterey, CA: ARI Field Unit at Presidio Monterey.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition, 15*, p. 445.

Gleick, J. (1987). *Chaos: Making a new science*. NY: Viking.

Gorman, R.P. and Sejnowski, T.J. (1988a). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, *1*, 75-89.

Gorman, R.P. and Sejnowski, T.J. (1988b). Learned classification of sonar targets using a massively parallel network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(7), 1135-1140.

Hinton, J.W. and Sejnowski, T.J. (1986). Learning and relearning in Boltzmann Machines. In Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, *Parallel Distributed Processing, Vol. I*, Chapter 7, Cambridge, MA.

Holland, J.H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79*, 2554-2558.

Kahan, J.P. et al. (1985). *Individual characteristics and unit performance* (R-3194-MIL). Santa Monica, CA: Rand Corporation.

Laskey, K.B. (in preparation). Adapting connectionist learning to Bayes Networks. Under revision for *International Journal of Approximate Reasoning*.

McClelland, J.L. and Elman, J.L. (1986). Interactive processes in speech perception: The TRACE model. In McClelland, J.L., Rumelhart, D.E., and the PDP Research Group, *Parallel Distributed Processing, Vol. II*, Cambridge, MA: MIT Press.

McClelland, J.L. and Rumelhart, D.E. (1988). *Explorations in parallel distributed processing*. Cambridge, MA: MIT Press.

Reilly, D.L., Cooper, L.N. and Elbaum, C. (1982). A neural model for category learning. *Biological Cybernetics*, *45*, 35-41.

Rosenblatt, F. (1958). Two theorems of statistical separability in the perceptron. In *Mechanization of thought processes: Proceedings of a symposium held at the National Physical Laboratory, Vol. 1*, 421-456. London: HM Stationery Office.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, *Parallel Distributed Processing, Vol. I*, Cambridge, MA: MIT Press.

Rumelhart, D.E., and Zipser, D. (1986). Feature discovery by competitive learning. In Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, *Parallel Distributed Processing, Vol. I*, Cambridge, MA: MIT Press.

Shastri, L. (1989). Default reasoning in semantic networks: A formalization of recognition and inheritance. *Artificial Intelligence*, *39*(3), 283-356.

Shisko, R. and Paulson, R.M. (1981). *Relating resources to the readiness and sustainability of combined arms units* (R-2769-MRAL). Santa Monica, CA: Rand.

The AVALON HILL Game Company (1975). TOBRUK. Tank Battles in North Africa: 1942.

Word, COL L.E. (1987). *Observations from three years at the National Training Center*, Research Product 87-02. Monterey, CA: ARI Field Unit, A Presidio of Monterey.

APPENDIX:

RULES FOR MANUAL WARGAME

Additional Rules for TOBRUK:

1. *Smoke (Obscuration).* Smoke in the line of sight between the firer and the engaged vehicle obscures the engaged vehicle from being acquired.

2. *Mines (Obstacles).* No more than four obstacles may be placed next to each other.

3. *Victory condition.* The attacker must move one of his vehicles onto the defender's "base," which is a single hex located at I22 on the gameboard.

4. *Combat - Firepower.* The British Grant tanks were only allowed to fire their 37mm gun and their 75mm gun was treated as if it did not exist.

5. *Combat - Weapons.* The Germans were assumed to be using all APCR (armor piercing) rounds. This gave them a greater range in which they could fire and higher probability of damaging or destroying the defending British tanks.

6. *Stacking.* The British were allowed a maximum of 3 Grants in a hex and the Germans were allowed a maximum of 4 Pzkw-IIIh's in a hex and 8 Italian M13/40's.

NOTE: To update TOBRUK to reflect the characteristics of present day conventional weapons would be a simple process. With current knowledge of the speed, armor, accuracy, lethality, and range of present day tanks, new movement factors, and hit probability and damage assessment tables could be made. Also the size of hexs could all be made relative to the speed of today's tanks to keep the battlefield in scale. Even such changes in technology that now allow tanks through to see through smoke or electronic jamming devices that block or confuse firing systems of opponents could be accounted for fairly easily.